Assignment 3: Data Exploration

Rebecca Marx

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics (ENV872L) on data exploration.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Use the lesson as a guide. It contains code that can be modified to complete the assignment.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file. You will need to have the correct software installed to do this (see Software Installation Guide) Press the `Knit` button in the RStudio scripting panel. This will save the PDF output in your Assignments folder.
6. After Knitting, please submit the completed exercise (PDF file) to the dropbox in Sakai. Please add your last name into the file name (e.g., "Salk_A02_DataExploration.pdf") prior to submission.

The completed exercise is due on Thursday, 31 January, 2019 before class begins.

1) Set up your R session

Check your working directory, load necessary packages (tidyverse), and upload the North Temperate Lakes long term monitoring dataset for the light, temperature, and oxygen data for three lakes (file name: NTL-LTER_Lake_ChemistryPhysics_Raw.csv). Type your code into the R chunk below.

```
getwd()
```

```
## [1] "C:/Users/rsmar/OneDrive/Documents/Spring 2019/RFolder/Environmental_Data_Analytics/Assignments"
```

```
#Load package
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------- tidyverse 1.2.1 -
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
## -- Conflicts -------------------------------------------------------------- tidyverse_conflicts() -
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#Import datasets
North.Temperate.Lakes.data <-read.csv("../Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")

#Looked at ReadMe file
```

2) Learn about your system

Read about your dataset in the NTL-LTER README file. What are three salient pieces of information you gained from reading this file?

ANSWER: 1) From the README file I learned that this data set was prepared specifically for this course at Duke University by Kateri Salk. 2) The data pertains to lakes in Wisconsin and it was collected as part of a Long Term Ecological Research study between 1984 and 2016. 3) The file provides information about methods for taking samples. For example, I leaned that physical and chemical variables were measured at the deepest points of each lake. Samples were sent to the Cary Institute of Ecosystem studies for analysis.

3) Obtain basic summaries of your data

Write R commands to display the following information:

1. dimensions of the dataset
2. class of the dataset
3. first 8 rows of the dataset
4. class of the variables lakename, sampledate, depth, and temperature
5. summary of lakename, depth, and temperature

```
# 1
dim(North.Temperate.Lakes.data)
```

```
## [1] 38614    11
```

```
# 2
class(North.Temperate.Lakes.data)
```

```
## [1] "data.frame"
```

```
# 3
head(North.Temperate.Lakes.data, 8)
```

```
##   lakeid  lakename year4 daynum sampledate depth temperature_C
## 1      L Paul Lake  1984    148    5/27/84  0.00          14.5
## 2      L Paul Lake  1984    148    5/27/84  0.25            NA
## 3      L Paul Lake  1984    148    5/27/84  0.50            NA
## 4      L Paul Lake  1984    148    5/27/84  0.75            NA
## 5      L Paul Lake  1984    148    5/27/84  1.00          14.5
## 6      L Paul Lake  1984    148    5/27/84  1.50            NA
## 7      L Paul Lake  1984    148    5/27/84  2.00          14.2
## 8      L Paul Lake  1984    148    5/27/84  3.00          11.0
##   dissolvedOxygen irradianceWater irradianceDeck comments
## 1             9.5            1750           1620     <NA>
## 2              NA            1550           1620     <NA>
## 3              NA            1150           1620     <NA>
## 4              NA             975           1620     <NA>
## 5             8.8             870           1620     <NA>
## 6              NA             610           1620     <NA>
## 7             8.6             420           1620     <NA>
## 8            11.5             220           1620     <NA>
```

```
# 4
class(North.Temperate.Lakes.data$lakename)
```

```
## [1] "factor"
```

```
class(North.Temperate.Lakes.data$sampledate)
```

```
## [1] "factor"
```

```
class(North.Temperate.Lakes.data$depth)
```

```
## [1] "numeric"
```

```
class(North.Temperate.Lakes.data$temperature_C)
```

```
## [1] "numeric"
```

```
# 5
summary(North.Temperate.Lakes.data$lakename)
```

```
## Central Long Lake     Crampton Lake    East Long Lake  Hummingbird Lake
##               539              1234              3905               430
##         Paul Lake        Peter Lake      Tuesday Lake         Ward Lake
##             10325             11288              6107               598
##    West Long Lake
##              4188
```

```
summary(North.Temperate.Lakes.data$depth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.50    4.00    4.39    6.50   20.00
```

```
summary(North.Temperate.Lakes.data$temperature_C)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.30    5.30    9.30   11.81   18.70   34.10    3858
```

Change sampledate to class = date. After doing this, write an R command to display that the class of sammpledate is indeed date. Write another R command to show the first 10 rows of the date column.

```
North.Temperate.Lakes.data$sampledate <- as.Date(North.Temperate.Lakes.data$sampledate, format = "%m/%d/

class(North.Temperate.Lakes.data$sampledate)
```

```
## [1] "Date"
```

```
head(North.Temperate.Lakes.data$sampledate, 10)
```

```
##  [1] "1984-05-27" "1984-05-27" "1984-05-27" "1984-05-27" "1984-05-27"
##  [6] "1984-05-27" "1984-05-27" "1984-05-27" "1984-05-27" "1984-05-27"
```

Question: Do you want to remove NAs from this dataset? Why or why not?

> ANSWER: No, I do not want to remove NAs from this dataset. Removing NAs from this data set would remove whole rows of information, even if there are data points in some of the comlumns. For example, there are several rows that have an entry for temperature, but do not have information for irradiance water or irradiance deck. Removing the NAs would remove some of the temperature entries. I am interested in summarizing all the temperature data that was avalailable. Since I am simply summarizing data and not doing anything like a regression, the NAs will not be problematic in my analysis. In fact, the graphs will automatically account for the missing data by "[removing] row containing non-finitie values".

4) Explore your data graphically

Write R commands to display graphs depicting:

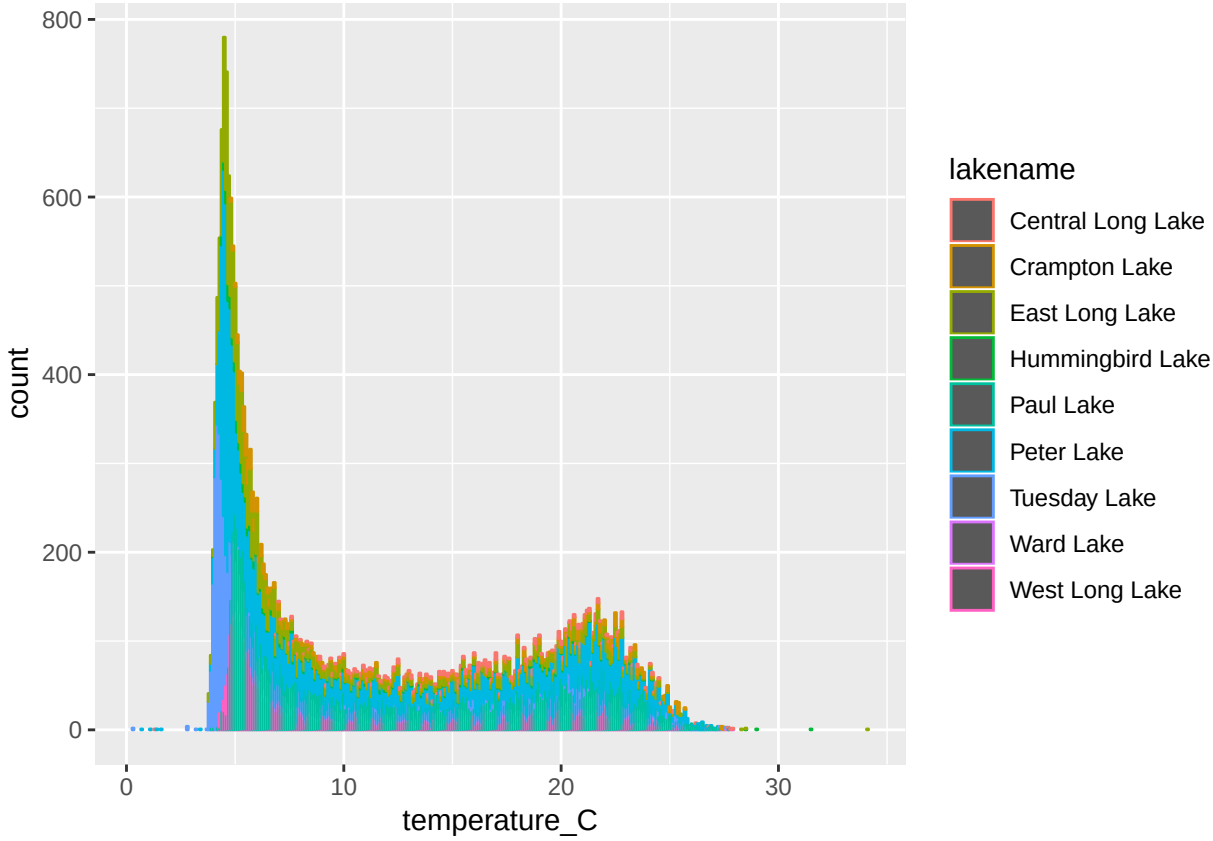
1. Bar chart of temperature counts for each lake
2. Histogram of count distributions of temperature (all temp measurements together)
3. Change histogram from 2 to have a different number or width of bins

4. Frequency polygon of temperature for each lake. Choose different colors for each lake.
5. Boxplot of temperature for each lake
6. Boxplot of temperature based on depth, with depth divided into 0.25 m increments
7. Scatterplot of temperature by depth

```
# 1
ggplot(North.Temperate.Lakes.data, aes(temperature_C, colour = lakename)) + geom_bar()
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_count).
```

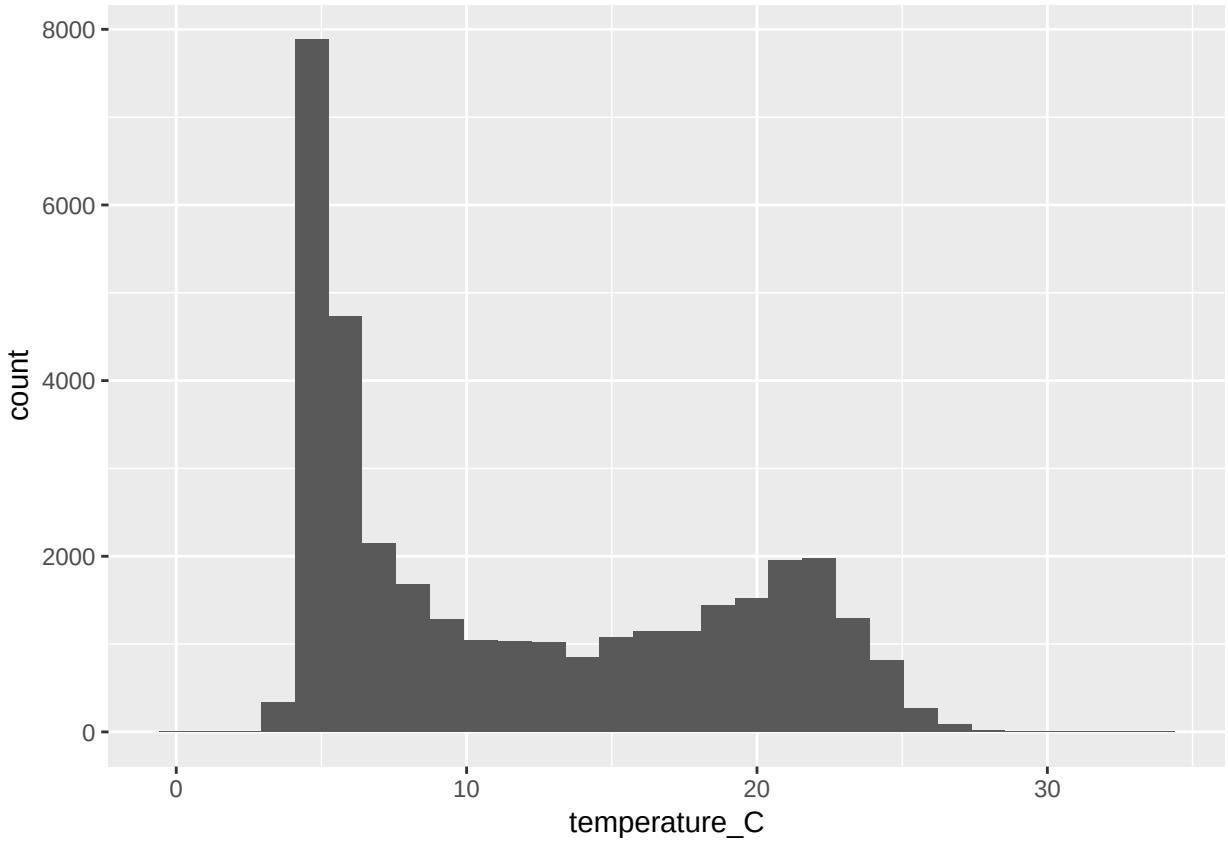
```
## Warning: position_stack requires non-overlapping x intervals
```




```
# 2
ggplot(North.Temperate.Lakes.data) + geom_histogram(aes(x = temperature_C))
```

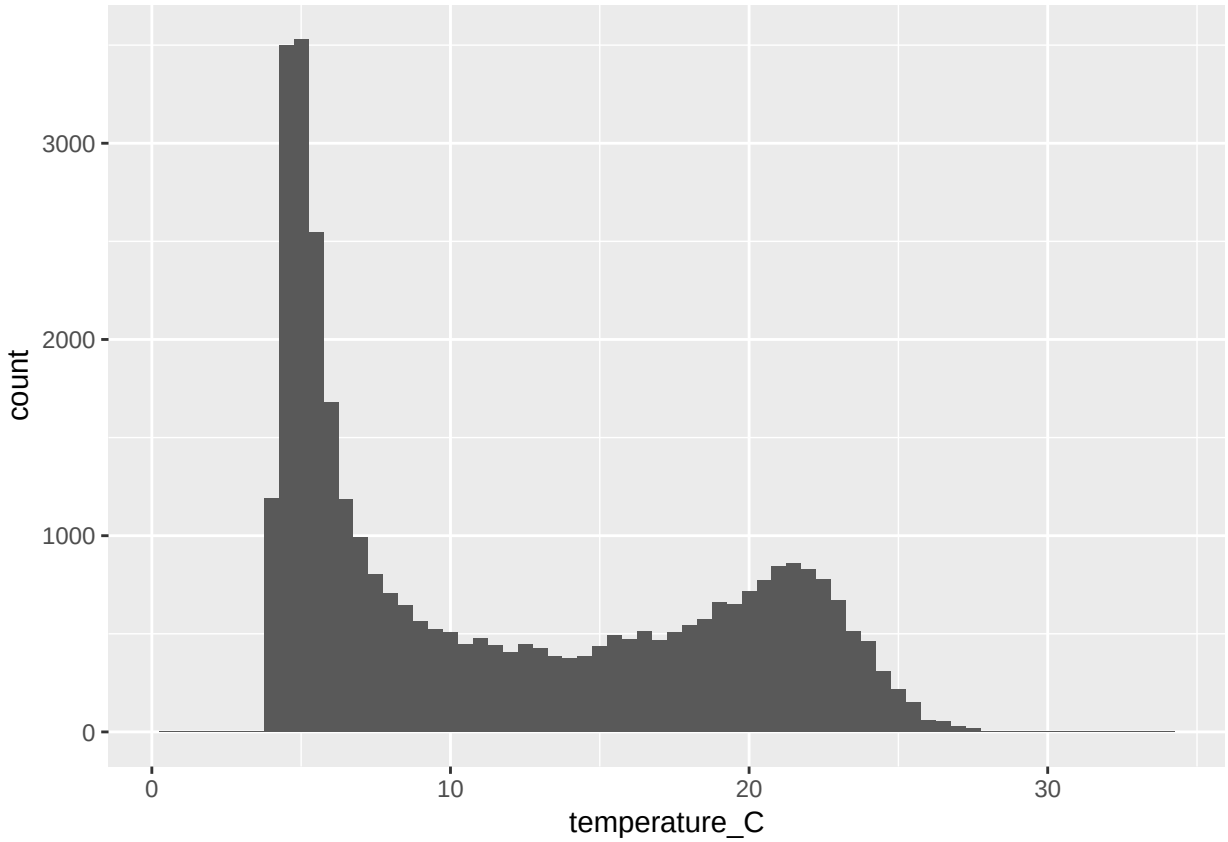
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```



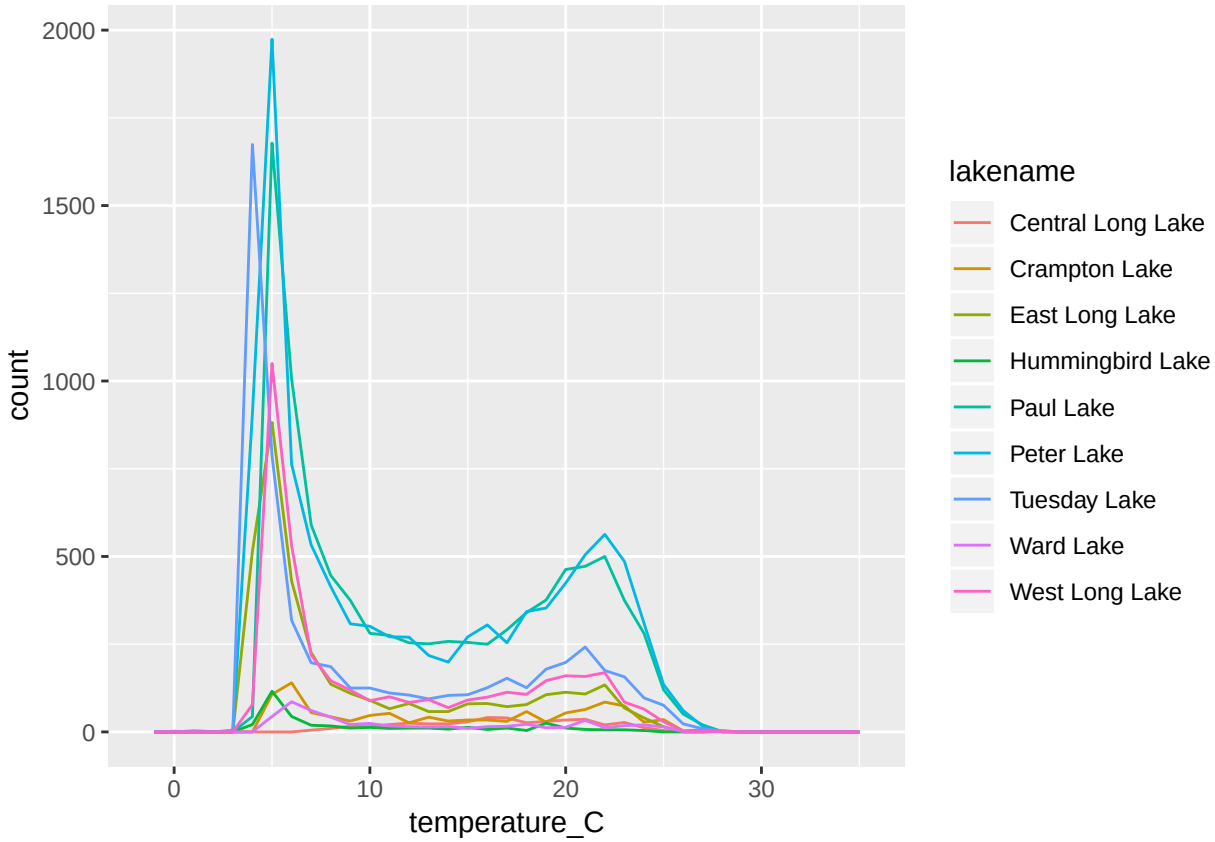

```
# 3
ggplot(North.Temperate.Lakes.data) + geom_histogram(aes(x = temperature_C), binwidth = .5)
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```



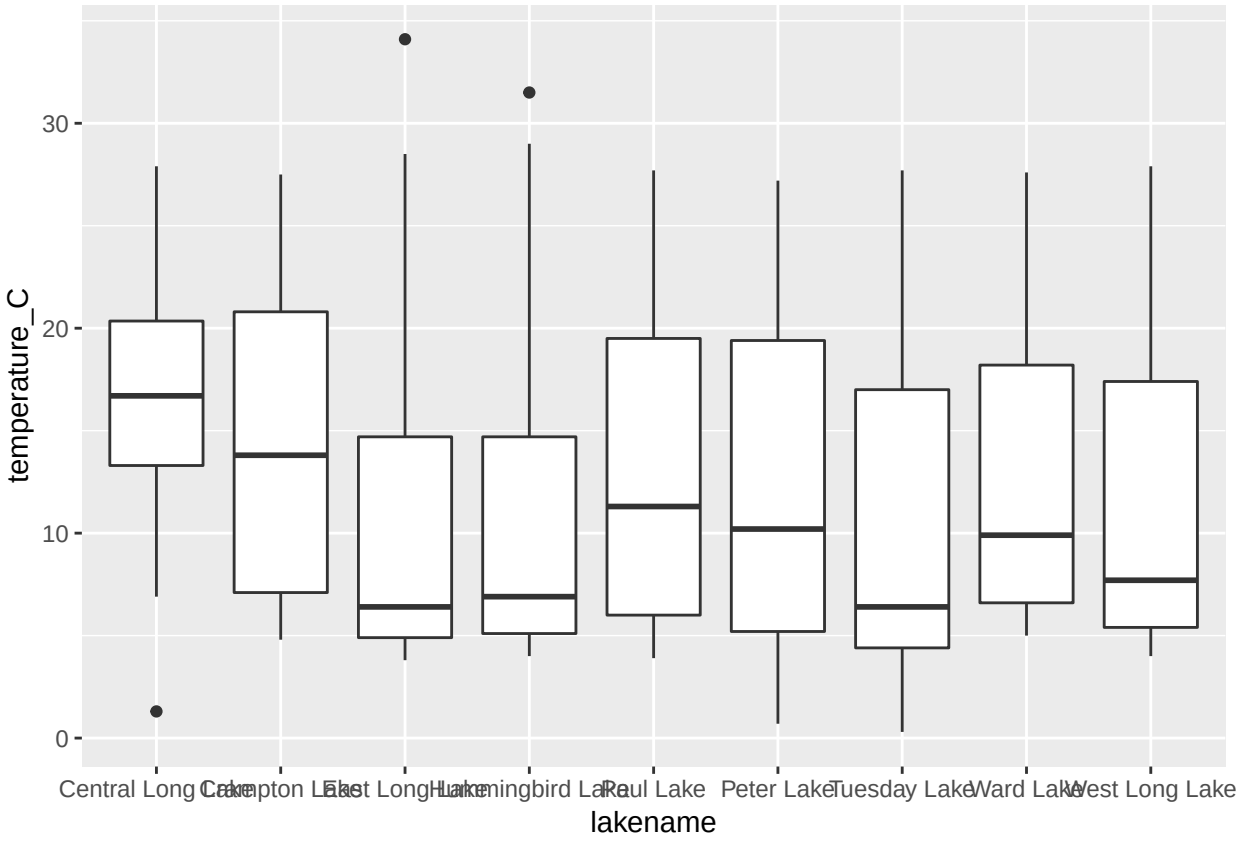

```
# 4
ggplot(North.Temperate.Lakes.data, aes(temperature_C, colour = lakename)) + geom_freqpoly (binwidth = 1
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_bin).
```



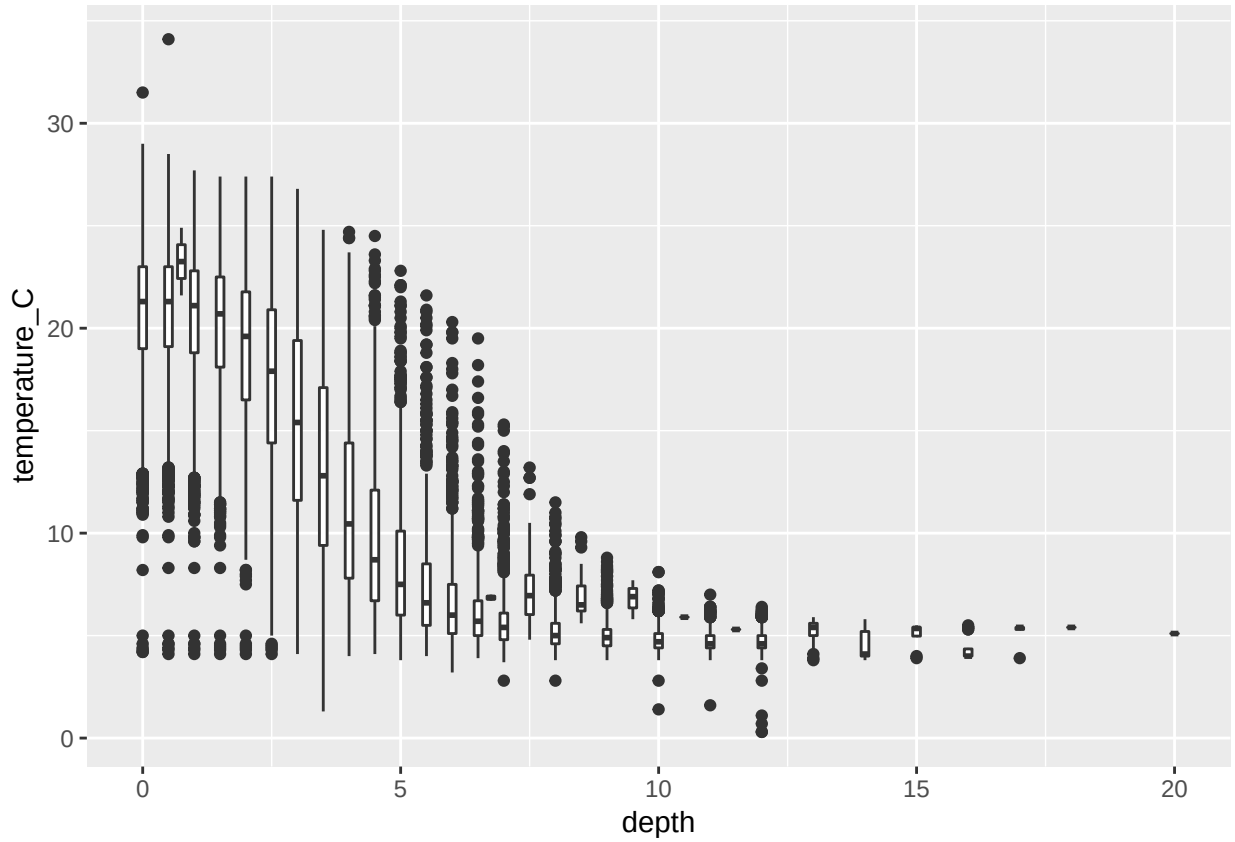

```
# 5
ggplot(North.Temperate.Lakes.data, aes(x = lakename, y = temperature_C)) + geom_boxplot()
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_boxplot).
```



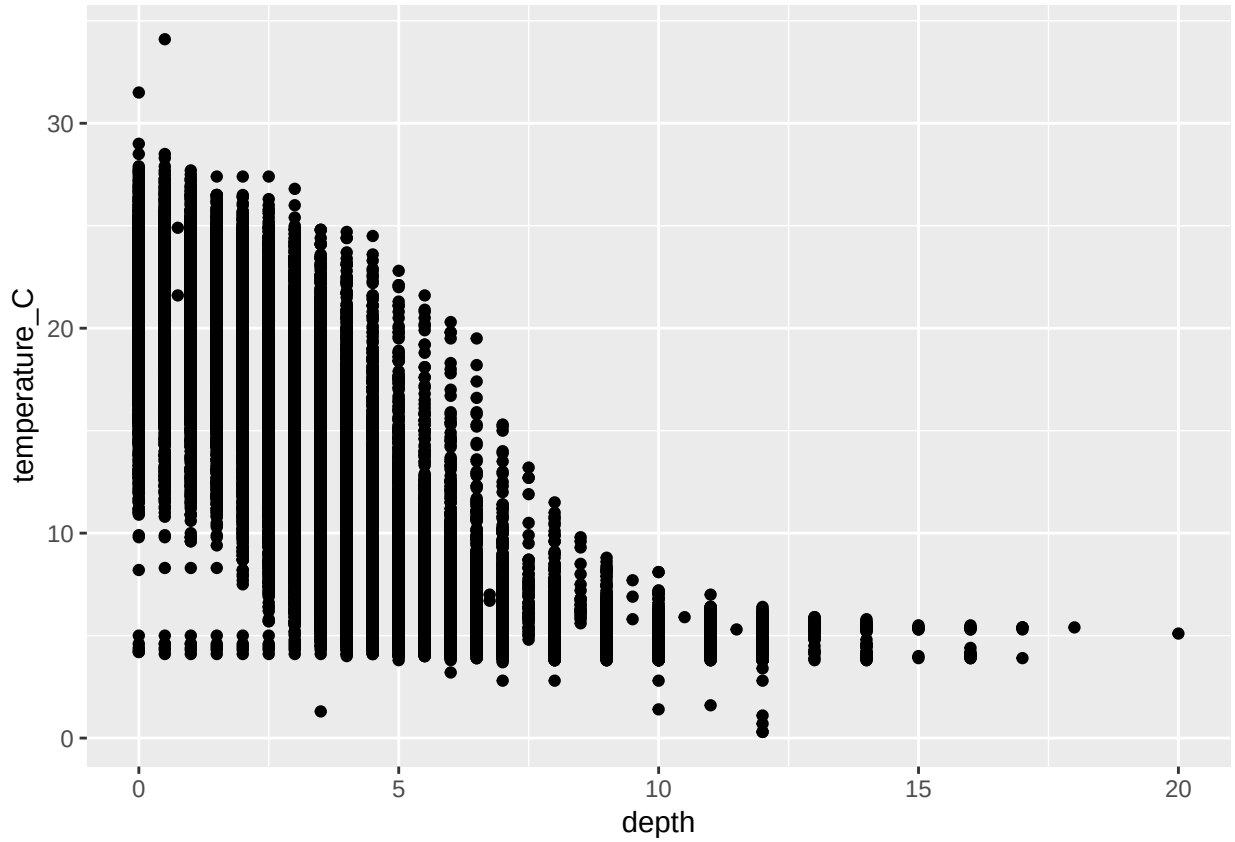

```
# 6
ggplot(North.Temperate.Lakes.data) + geom_boxplot(aes(x = depth, y = temperature_C, group = cut_width(d
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_boxplot).
```




```
# 7
ggplot(North.Temperate.Lakes.data) + geom_point(aes(x = depth, y = temperature_C))
```

```
## Warning: Removed 3858 rows containing missing values (geom_point).
```



5) Form questions for further data analysis

What did you find out about your data from the basic summaries and graphs you made? Describe in 4-6 sentences.

> ANSWER: From my basic summaires, I found out that Peter Lake has the most temperature counts, and Hummingbird Lake has the least. The highest frequnecy of temperature is around 5 degrees celcius. Temperatures tend to be higher at lower depths. Central Long Lake has the highest median temperature. There were very few outliers, but the higest outlier temperature was close to 35 degress Celcius in East Long Lake while the lowest was close to 0 degress Celcius in Central Long Lake.

What are 3 further questions you might ask as you move forward with analysis of this dataset?

> ANSWER 1: How has the temperature in each lake changed over time?
>
> ANSWER 2: How does water temperature relate to levels of dissolved oxygen?
>
> ANSWER 3: What is the trend in temperature in terms of months of the year?